Detección de regiones del Uruguay climáticamente homogéneas respecto de sus vientos extremos

Denis, Andrés [1], *Durañona, Valeria* [1], *Molinari, Mariana*[1]
Perera, Gonzalo[2], *Piccini, Juan*[3]

[1] * IMFIA, FING, UdelaR
[2] MEDIA, CURE, UdelaR
[3] IMERL, FING, UdelaR
*valeriad@fing.edu.uy

Contents

1 Introducción

Los eventos de vientos extremos provocan graves daños a estructuras tales como edificios, casas, silos, muelles, galpones, invernáculos, tendidos eléctricos, etc.

Estos eventos pueden clasificarse en Sinópticos y No Sinópticos, con características físicas y estadísticas diferentes:

- Escala espacial (dimensiones de la zona afectada simultáneamente por el evento).
- Escala temporal (tiempo durante el cual se sostienen velocidades intensas).
- Perfil de velocidad en altura.

- En nuestro país los datos muestran que los eventos más fuertes son mayoritariamente del tipo No Sinópticos.
- La importancia relativa de los eventos sinópticos sobre los no sinópticos crece hacia el sureste del país.
- Rocha sería una de las zonas del país donde los eventos sinópticos predominan más que en otras zonas.
- En el norte (por ej. Salto, Paysandú) los eventos no sinópticos tienden a ser más intensos y frecuentes que en otras zonas del país.

2 Metodología

Con los máximos diezminutales de la velocidad de ráfaga se construye un valor de referencia, que es el valor a 45 mt. de altura, interpolado a partir de los dos valores de velocidad más cercanos en altura.

Cada evento recoge una ventana de diez horas de ancho centrada en el valor máximo registrado, un total de 61 mediciones diezminutales.

Estas 61 mediciones configuran un vector que al graficarse nos da lo que llamaremos curva de velocidad de ráfaga diezminutal a 45 mt de altura, o curva de velocidad a 45 mt de altura.

Previamente se utilizó una muestra de 251 eventos clasificada por expertos en sinópticos y no sinópticos para entrenar algoritmos de clasificación. La variable utilizada para ello fue precisamente la curva de velocidad a 45 mt de altura.

Mediante la Dynamic Time Warping o DTW (que es una medida de similaridad que permite comparar la forma entre distintas curvas y asignar una suerte de distancia entre curvas) se construyó una matriz de "distancias "(similitudes) entre las 251 curvas ya etiquetadas.

Luego se implementaron algoritmos de clasificación y clusterización que tomaban como input la matriz de similitudes, agrupando los eventos en función del parecido entre las curvas de velocidad asociadas a cada evento.

La hipótesis de trabajo es que los eventos no sinópticos tienen curvas de velocidad con una forma característica, distinta a la de los eventos sinópticos.

Se mide el grado de homogeneidad de los grupos así formados, siendo de esperar que eventos del mismo tipo sean asignados por los algoritmos al mismo grupo.

Una vez confirmado que la clasificación automática de eventos mediante dichas curvas de velocidad tiene altas tasas de acierto (superiores al 90%), se procedió a la detección automática de eventos en cada uno de los sitios, trabajando con la velocidad a 45 mt de altura (máximos diezminutales).

Fijado un piso de 22.2 mt/s (80 Km/h, velocidad a partir de la cual los daños causados por el viento comienzan a notarse) procedemos como sigue:

1. Localizamos el máximo y lo extraemos junto con las 30 lecturas anteriores y posteriores al mismo, teniendo el registro de las 5 horas anteriores y posteriores a dicho máximo. Este vector de 61 componentes es lo que denominamos "evento ", y su representación gráfica es lo que hemos llamado curva de velocidad.
2. Repetimos el paso anterior con los datos remanentes hasta que no queden datos suficientes.

3 K vecinos más cercanos

Hecho esto, para cada sitio clasificaremos los eventos (o curvas de velocidad) así obtenidos como sinópticos (“s ”) o no sinópticos (“ns ”). La clasificación se hace mediante el método KNN (K Nearest Neighbors) de K vecinos más cercanos.

Para elegir el valor óptimo de K se dividió la muestra de 251 eventos clasificados por expertos en una muestra de entrenamiento y otra de testeo (conocemos las distancias entre cada par de sitios).

Esto se hizo con proporciones 2/3, 3/4, 4/5, 9/10 y 95/100, siendo la muestra de entrenamiento la mayor en todos los casos. Una vez partida la muestra en entrenamiento y testeo, para cada sitio se ordenan los restantes 250 por distancia creciente.

Luego para cada sitio de la muestra de test ordenamos sus vecinos de la muestra de entrenamientio según la distancia, de menor a mayor y vamos anotando las etiquetas del primer vecino más cercano, del segundo, etc.

Luego comparamos la etiqueta del sitio a testear con la de su primer vecino más cercano, el segundo, etc.

Para cada valor K=1,2,...,n (n es el tamaño de la muestra de entrenamiento) tomamos la etiqueta mayoritaria en los primeros K vecinos y comparamos con la etiqueta del sitio a testear.

Repetimos esto 20 veces para c/u de las proporciones utilizadas y elegimos el K que produjo una tasa de aciertos superior al 90%. En caso de existir varios K posibles elegimos al de mejor tasa de aciertos, resultando en K=3.

Hecho esto, clasificamos los eventos detectados automáticamente (para los cuales no hay etiqueta previa) hallando la distancia entre cada evento y los 251 eventos clasificados por expertos.

Usamos luego los 3 vecinos (eventos clasficados por expertos) más cercanos a cada sitio para asignarle la etiqueta mayoritaria entre dichos vecinos.

De esta manera a cada sitio le asociamos una palabra binaria, compuesta de los símbolos “ns”y “s”. Esta palabra puede convertirse también en una palabra de ceros y unos (1=ns, 0=s).

Por la forma en la que se extrajeron los eventos, el evento cuyo pico fue más alto es el que ocupa el primer lugar, seguido por el que tuvo en su máximo el segundo valor, etc.

Para cada sitio también guardamos el vector con los valores de velocidad de los picos de cada evento. Esto es, además de saber si un evento determinado fue clasificado como s o ns, tenemos también la velocidad máxima (pico) asociada a dicho evento.

Aquellos sitios donde los mayores eventos suelen ser “ns”tendrán en sus palabras a la mayoría de los símbolos “ns”.

4 Capturando la sinopticidad

¿Cómo capturar la proporción ns/s así como la forma en la que estos dos símbolos se distribuyen en cada palabra?

El indicador 1 (ind1) lo que hace es calcular el cociente acumulado ns/s. Esto es, dada una palabra asociada a un sitio, comenzando por el primer símbolo de la misma vamos calculando para los primeros k símbolos el cociente ns/s, con k=1,2,... hasta el largo de la palabra.

De este modo, un sitio donde los símbolos "ns"y "s"se alternen va a mostrar un id1 que comienza con 1 o 0 (según el primer símbolo sea "ns"o "s"respectivamente) y que luego oscilará en una banda de centro 0.5.

Un sitio donde los mayores eventos son "ns ", mostrará un ind1 que comienza en 1 y luego desciende en la medida que aparezcan rachas de eventos "s".

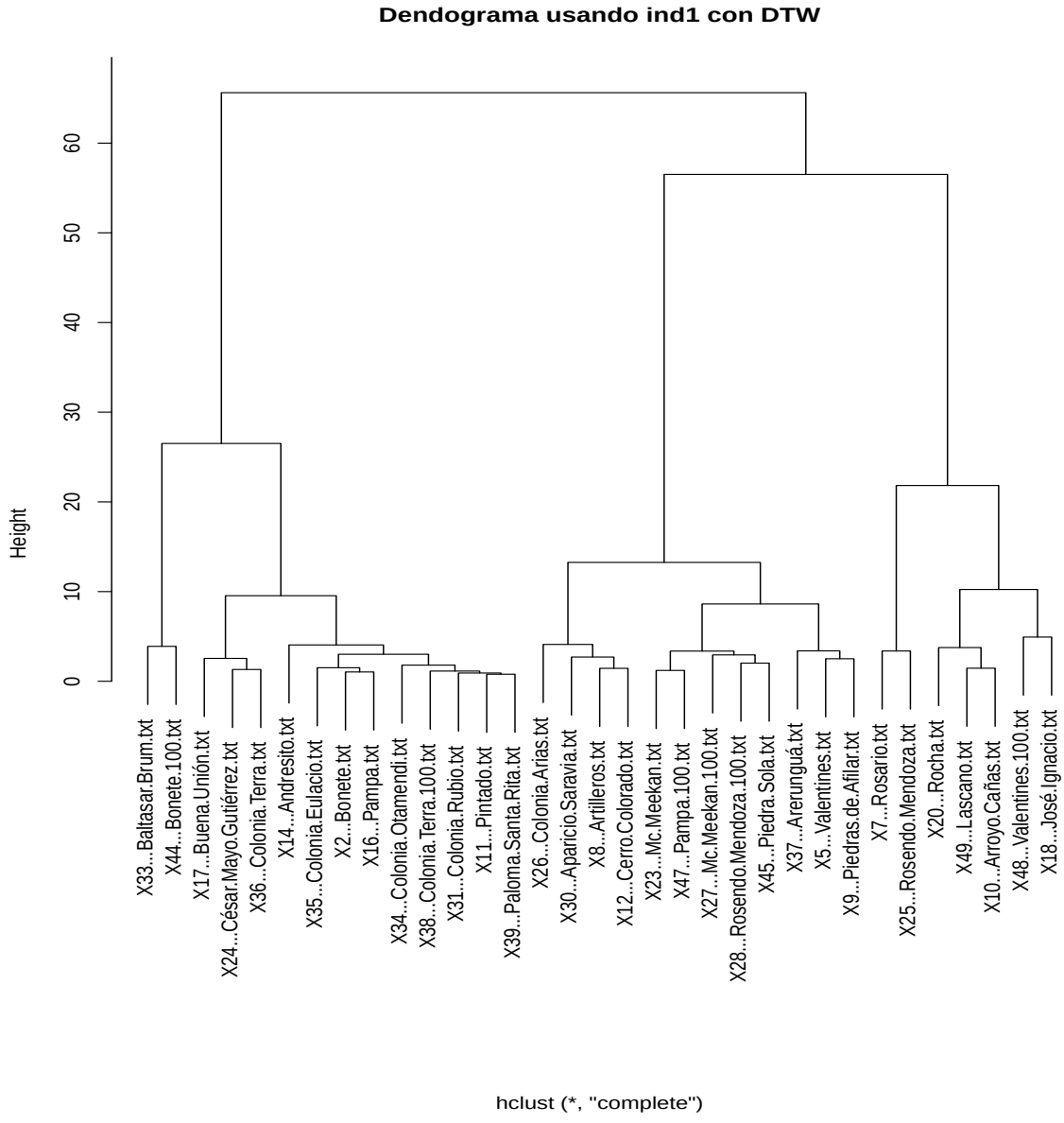
Si para cada sitio nos quedamos con los 10 primeros máximos (los mayores) y luego calculamos la proporción ns/s en los 34 sitios, obtenemos

0.882	0.794	0.794	0.765	0.706	0.706	0.853	0.706	0.647	0.588

Es claro que los eventos ns dominan en los 10 eventos mayores. A modo de ejemplo, si en cada sitio vemos el mayor evento registrado, en el 88.2% de dichos sitios dicho evento es no sinóptico.

5 Clustering para 45 mts. de altura

Calculamos la matriz de similaridades DTW entre las curvas ind1 de cada pareja de sitios y procedemos a clusterizar mediante métodos jerárquicos (Hclust). Usando linkage completo obtenemos el dendrograma de la figura (1).



..

Fig. 1: Dendograma jerárquico, 45 mts. de altura

Se aprecia la existencia de dos grandes grupos, uno de los cuales a su vez puede subdividirse en otros dos grupos. Las figuras (2) y (3) muestran los grupos que obtendríamos si consideramos dos o tres grupos respectivamente.

5.1 Grupos Robustos

Si aplicamos métodos no jerárquicos de clustering tales como PAM (Partition Around Medoids, que puede verse como una versión robusta del método de K-medias), podremos ver si hay sitios que sean asignados al mismo grupo por ambos métodos.

Partiendo en 2 y 3 grupos mediante PAM y comparamos con las correspondientes particiones obtenidas mediante Hclust, obtenemos las tablas (1) y (2).

	1	2
1	16	1
2	3	13

Tab. 1: Filas: Grupos PAM, Columnas: Grupos Hclust

	1	2	3
1	7	0	9
2	0	11	3
3	0	3	0

Tab. 2: Filas: Grupos PAM, Columnas: Grupos Hclust

5.2 2 Grupos Robustos

Cuando partimos en 2 grupos la tabla (1) muestra que tenemos 16 y 13 sitios respectivamente que son asignados juntos por ambos métodos. Nos referiremos a estos grupos como grupos robustos.

Los 16 sitios que ambos métodos asignan juntos (Grupo 1 robusto) son : 5 (Valentines), 7 (Rosario), 9 (Piedras de Afilar), 10 (Arroyo Cañas), 18 (José Ignacio), 20 (Rocha), 23 (Mc Meekan), 25 (Rosendo Mendoza), 26 (Colonia Arias), 27 (Mc Meekan 100), 28 (Rosario Mendoza 100), 37 (Arerunguá), 45 (Piedra Sola), 47 (Pampa 100), 48 (Valentines 100) y 49 (Lascano).

Con excepción de los sitios 37, 45 y 47, los restantes sitios están al sur del Río Negro.

Los 13 sitios que ambos métodos asignan juntos al Grupo 2 robusto son: 2 (Bonete), 11 (Pintado), 14 (Andresito), 16 (Pampa), 17 (Buena Unión), 24 (César Mayo Gutiérrez), 31 (Colonia Rubio), 33 (Baltasar Brum), 34 (Colonia Otamendi), 35 (Colonia Eulacio), 36 (Colonia Terra), 38 (Colonia Terra 100) y 39 (Paloma Santa Rita).

Excepto el sitio 11, los restantes se encuentran o bien sobre el Río Negro o al norte del mismo.

La figura (4) muestra estos dos grupos en el mapa.

Para cada sitio tenemos las velocidades de los picos de los eventos, si calculamos los cuartiles de dichos picos para cada sitio en c/u de los dos grupos obtenemos la figura (5). Para confirmar que la diferencia entre los dos cuartiles superiores es estadísticamente significativa, hacemos un test de permutaciones. Tomamos los grupos

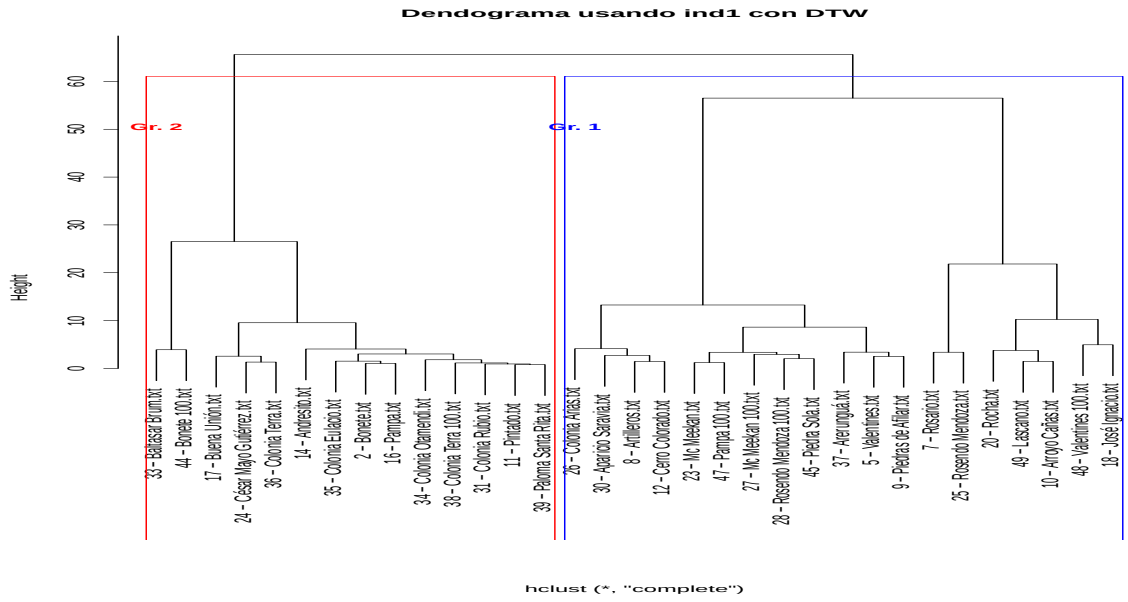


Fig. 2: Dendograma obtenido con Hclust. Grupo 1: 19 sitios, Grupo 2: 14 sitios

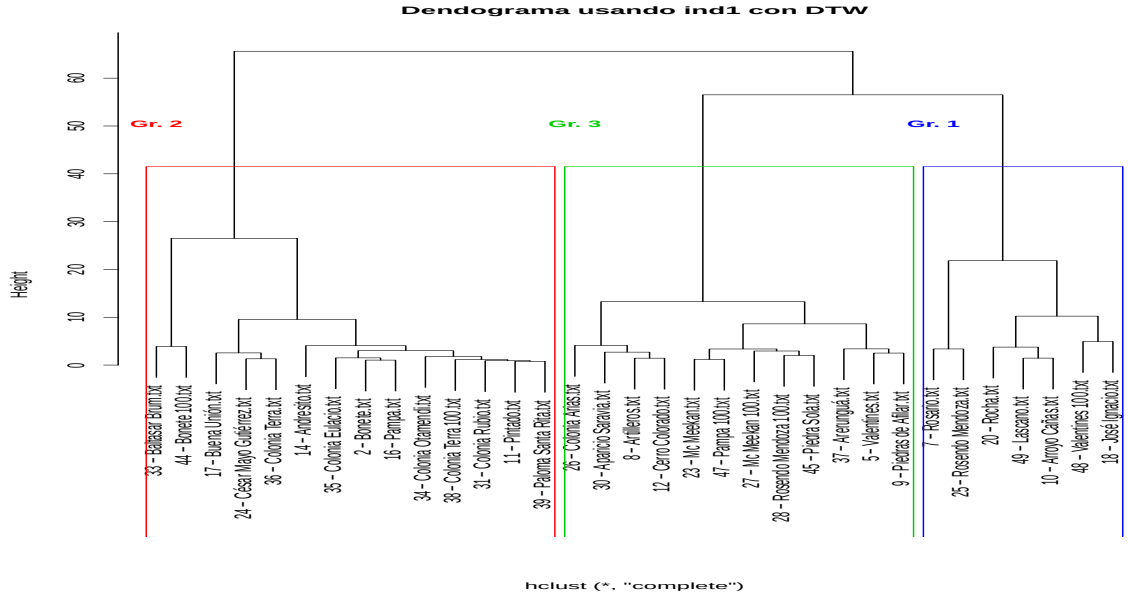


Fig. 3: Dendograma obtenido con Hclust. Grupo 1: 7 sitios, Grupo 2: 14 sitios, Grupo 3: 12 sitios

Grupos robustos 1 (Azul) y 2 (Rojo). Conformados por sitios que van juntos tanto por Hclust como por PAM cuando partimos en dos grupos por cada método. El Grupo 1 está mayoritariamente al sur del Río Negro. Los sitios sin colorear son asignados en forma cruzada por los dos métodos empleados.

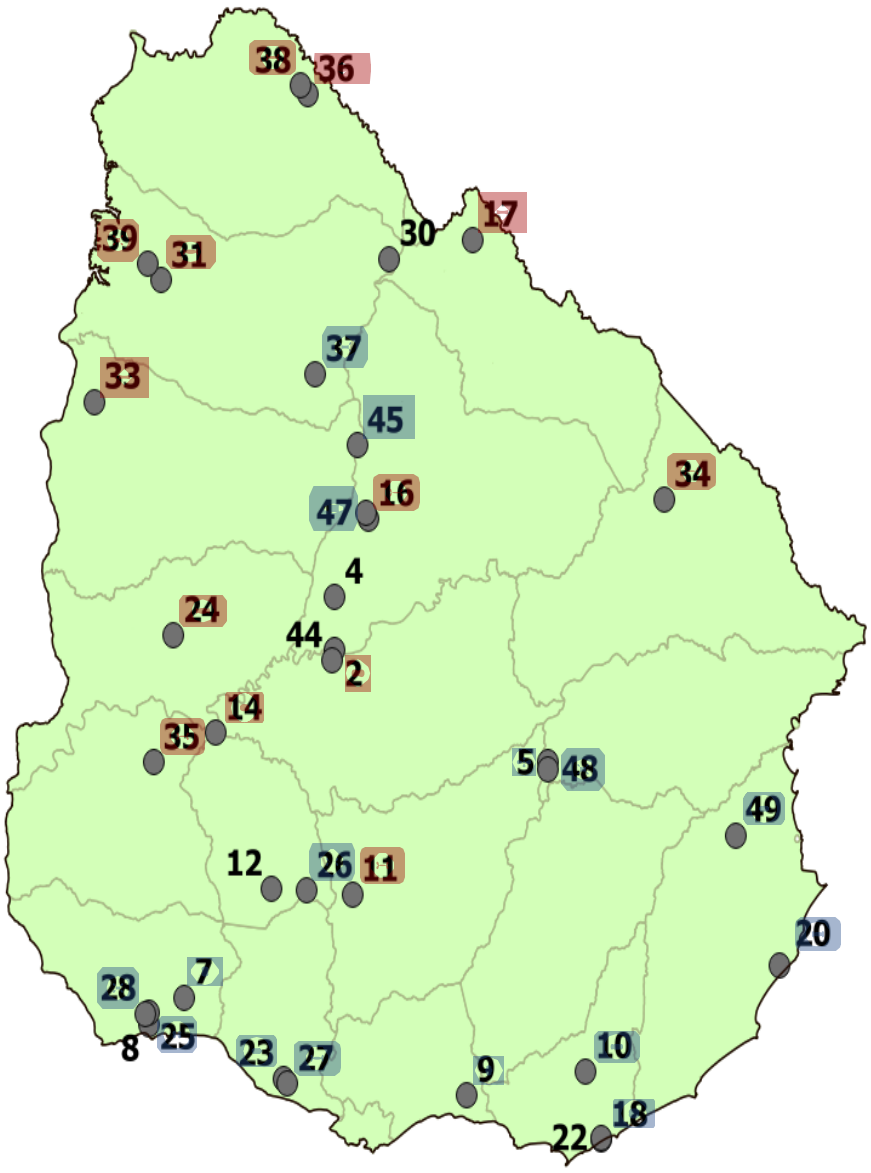
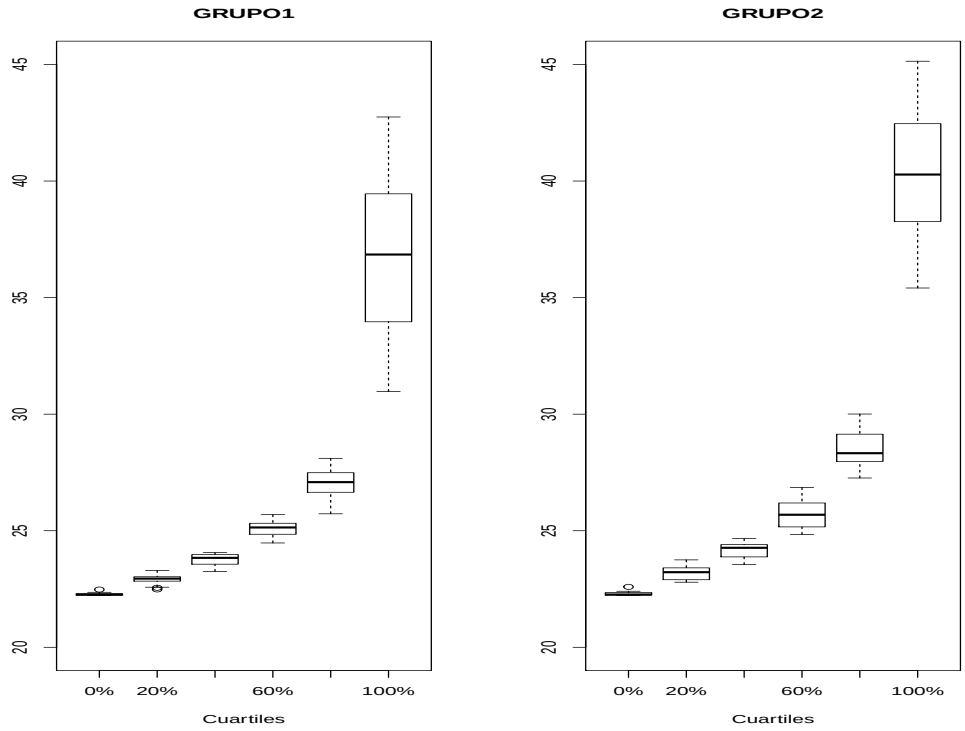


Fig. 4: Grupos resultantes de la tabla (1)



..

Fig. 5: Percentiles de los grupos robustos: Nótese la diferencia entre ambos grupos en los percentiles superiores

1 y 2, con 16 y 13 sitios respectivamente. Calculamos el promedio de cada cuartil en cada grupo y luego hacemos la diferencia entre dichos promedios.

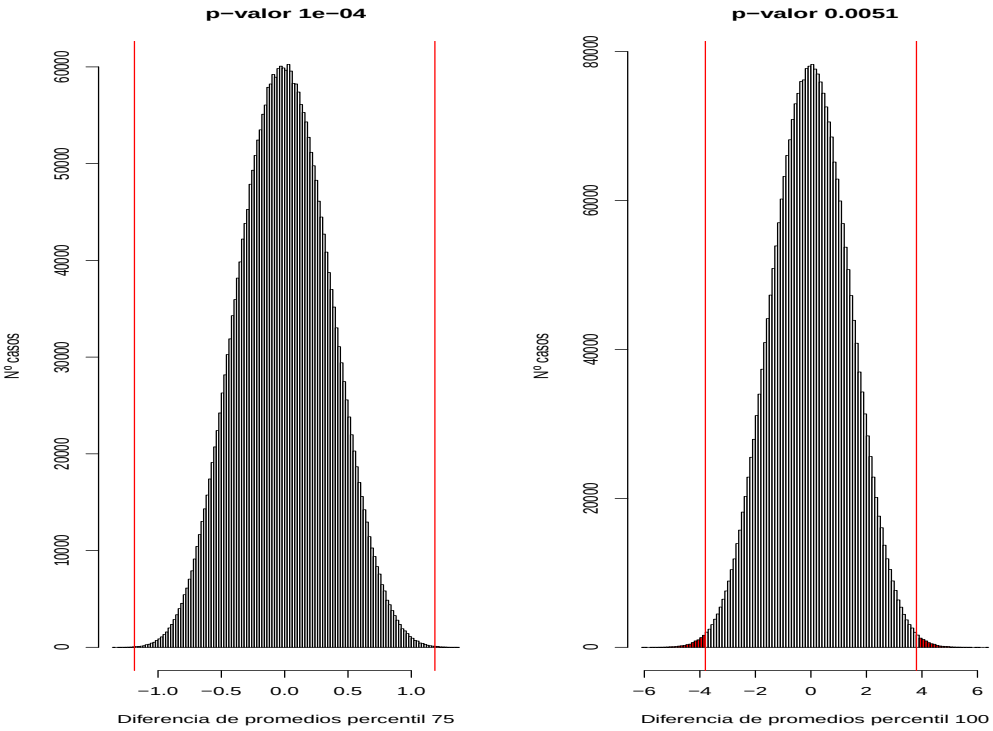
Luego separamos al azar los 29 sitios (16+13) en dos subconjuntos de 16 y 13 miembros respectivamente. Para c/u de las C_{16}^{29} formas posibles de hacer dicha separación, calculamos el promedio en los cuartiles de cada grupo y hacemos la diferencia de dichos promedios (distribución de aleatorización).

Las diferencias observadas en la muestra son -1.186254 y -3.802049 para los percentiles 75 y 100 respectivamente.

Si no hubiese diferencia entre los cuartiles de los grupos 1 y 2 entonces para cualquier asignación de 16 sitios a un grupo y 13 al otro deberíamos observar diferencias entre promedios similares.

Por tanto calculamos la probabilidad de obtener una diferencia tanto o más extrema que la observada, que será el p-valor del test. Podemos visualizarlo mediante el histograma de las diferencias obtenidas, marcando en rojo aquellas diferencias tanto o más extremas que la observada, como vemos en la figura (6).

Los p-valores obtenidos son muy cercanos a cero, lo que indica que las diferencias observadas son muy poco creíbles si suponemos que no hay diferencias entre los cuartiles de los grupos 1 y 2. Las figuras (7) y (8) muestran la variable ind1 en cada sitio.



..

Fig. 6: Distribución de aleatorización para la diferencia de promedios en los percentiles 75 y 100 de los grupos 1 y 2 de la figura (4)

Vemos que en el Grupo 2 todos los sitios muestran una mayoría de eventos ns, mientras que en el Grupo 1 la mitad SE de los sitios muestran un predominio de eventos s.

En dicho grupo, a medida que nos desplazamos hacia el W vemos que aumenta la proporción de eventos ns.

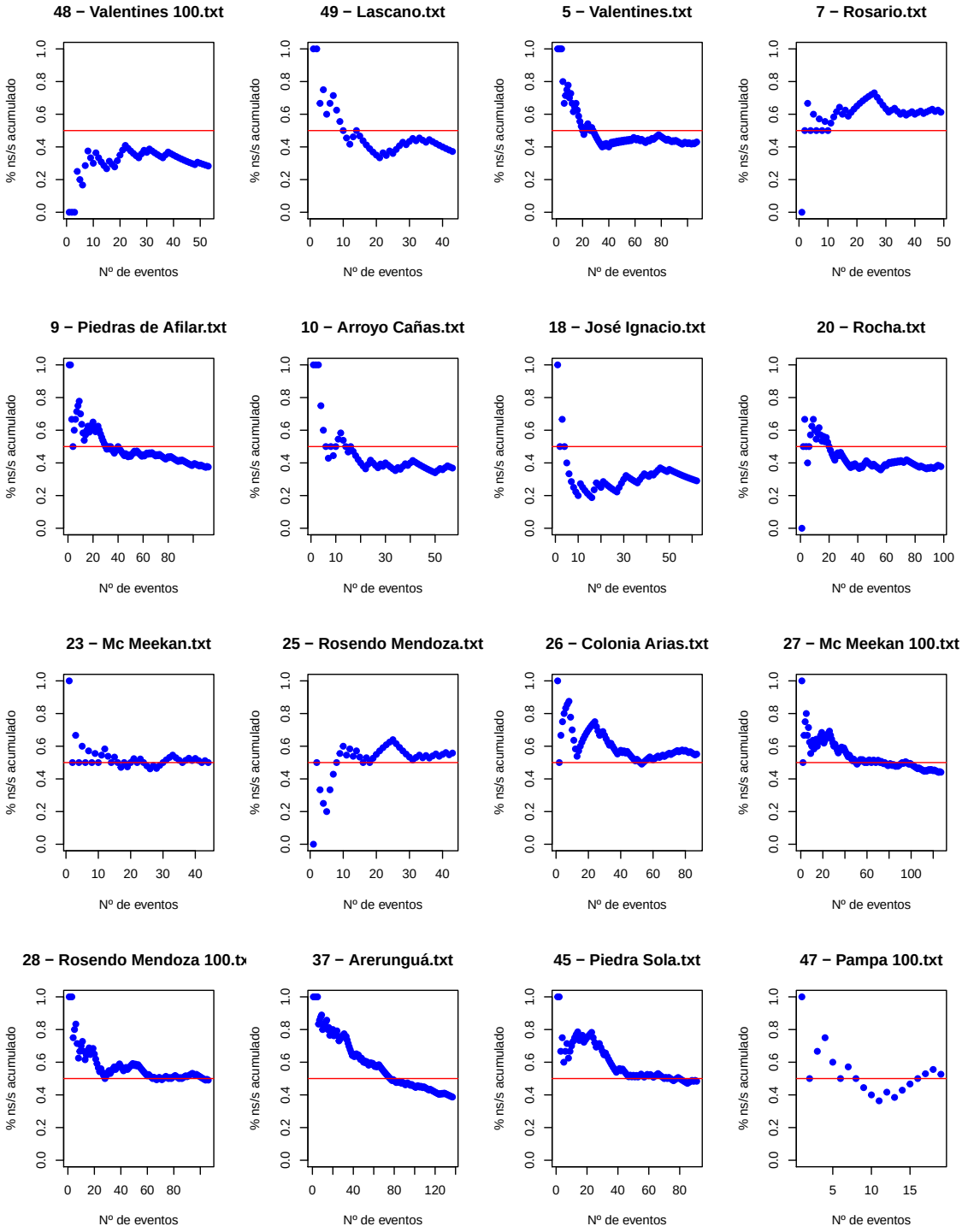


Fig. 7: Variable ind1 en el Grupo 1

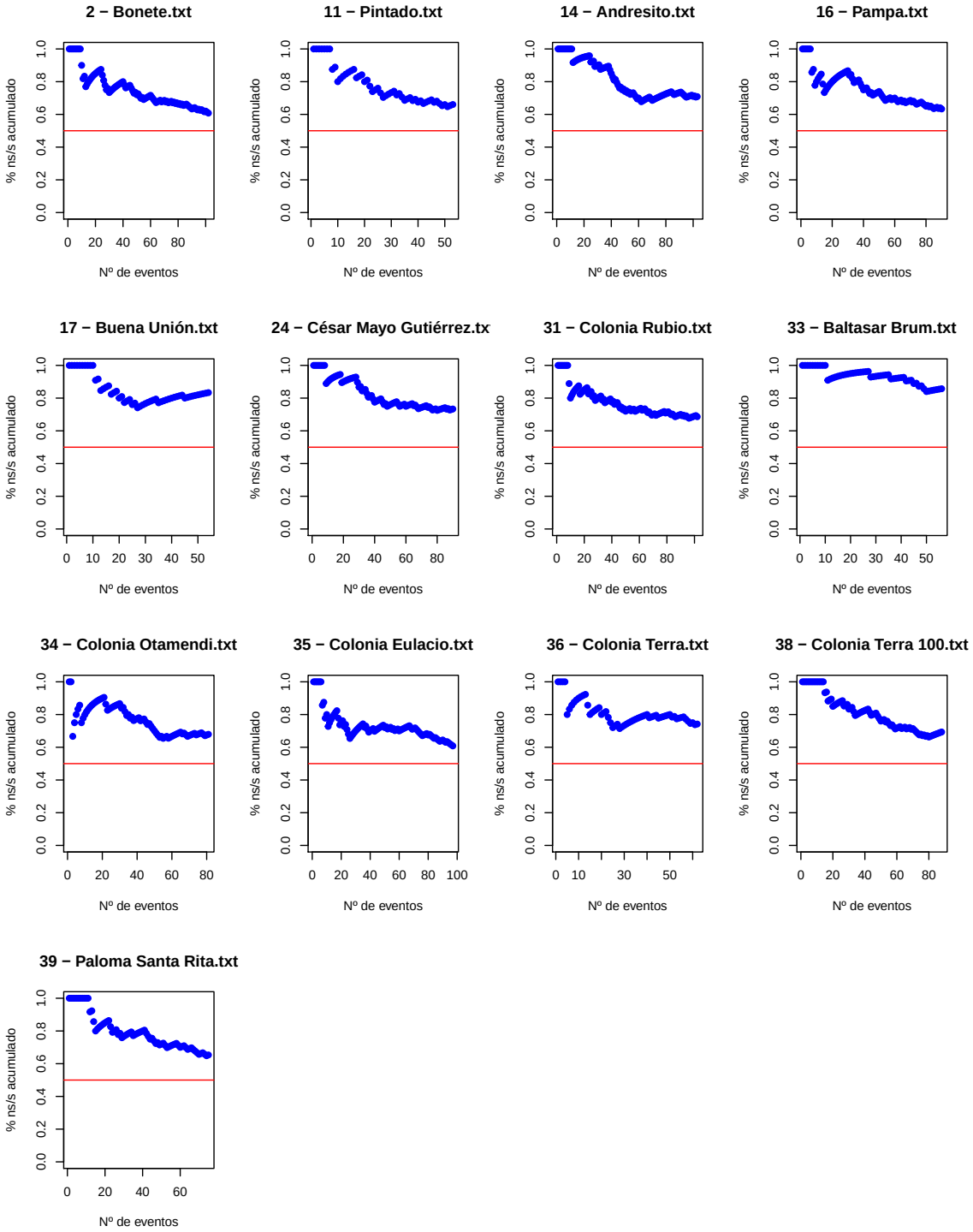


Fig. 8: Variable ind1 en el Grupo 2

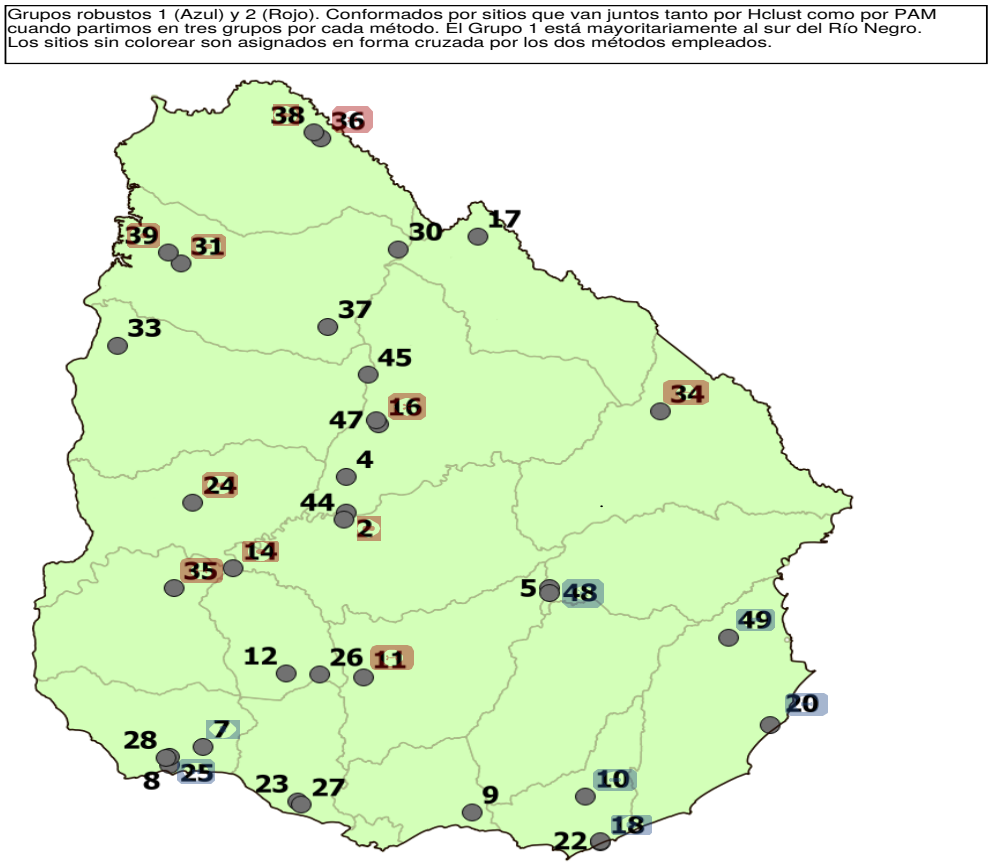
5.3 3 Grupos Robustos

Si vemos la tabla (2), cuando buscamos 3 clusters hay 7 sitios que tanto Hclust como PAM asignan al mismo grupo (Grupo 1) y 11 sitios que ambos métodos asignan al Grupo 2. No hay sitios que sean asignados por ambos métodos al Grupo 3.

Los 7 sitios del Grupo 1 son: 7 (Rosario), 10 (Arroyo Cañas), 18 (José Ignacio), 20 (Rocha), 25 (Rosendo Mendoza), 48 (Valentines 100) y 49 (Lascano). Son una parte del Grupo 1 Robusto obtenido cuando se separaba en dos clusters.

Los 11 sitios del Grupo 2 son: 2 (Bonete), 11 (Pintado), 14 (Andresito), 16 (Pampa), 24 (César Mayo Gutiérrez), 31 (Colonia Rubio), 34 (Colonia Otamendi), 35 (Colonia Eulacio), 36 (Colonia Terra), 38 (Colonia Terra 100) y 39 (Paloma Santa Rita), son una parte del Grupo 2 Robusto obtenido en la instancia anterior.

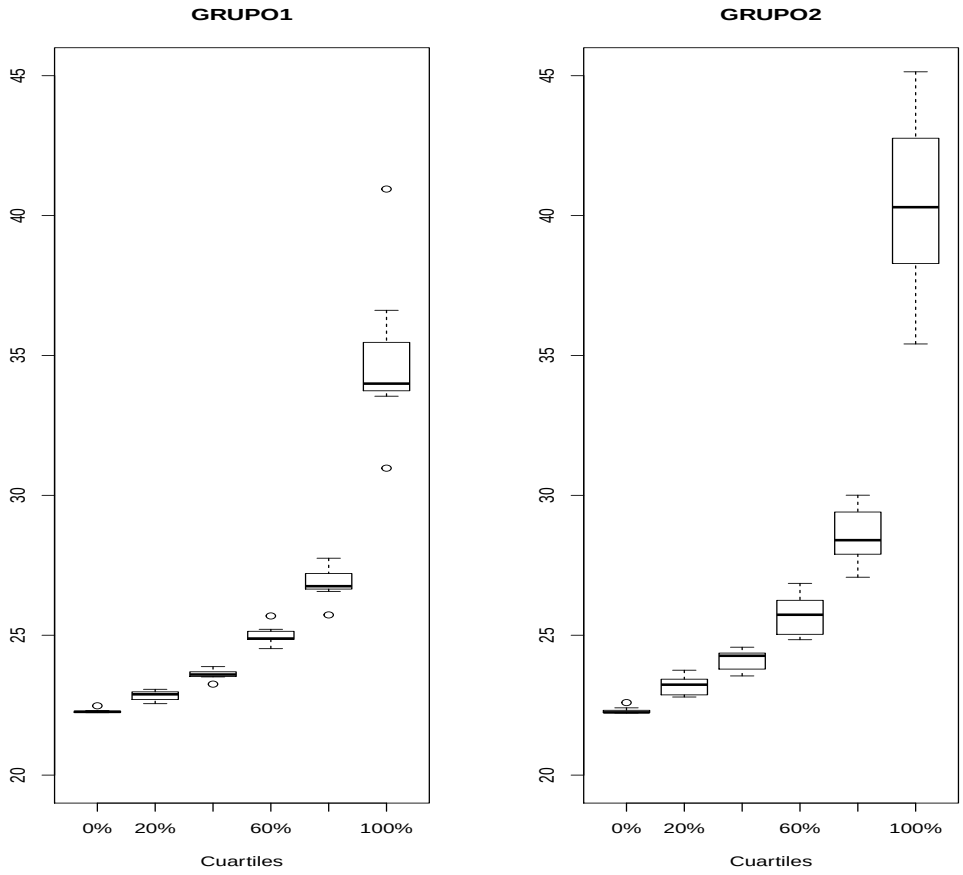
La figura (9) muestra estos dos grupos robustos.



..

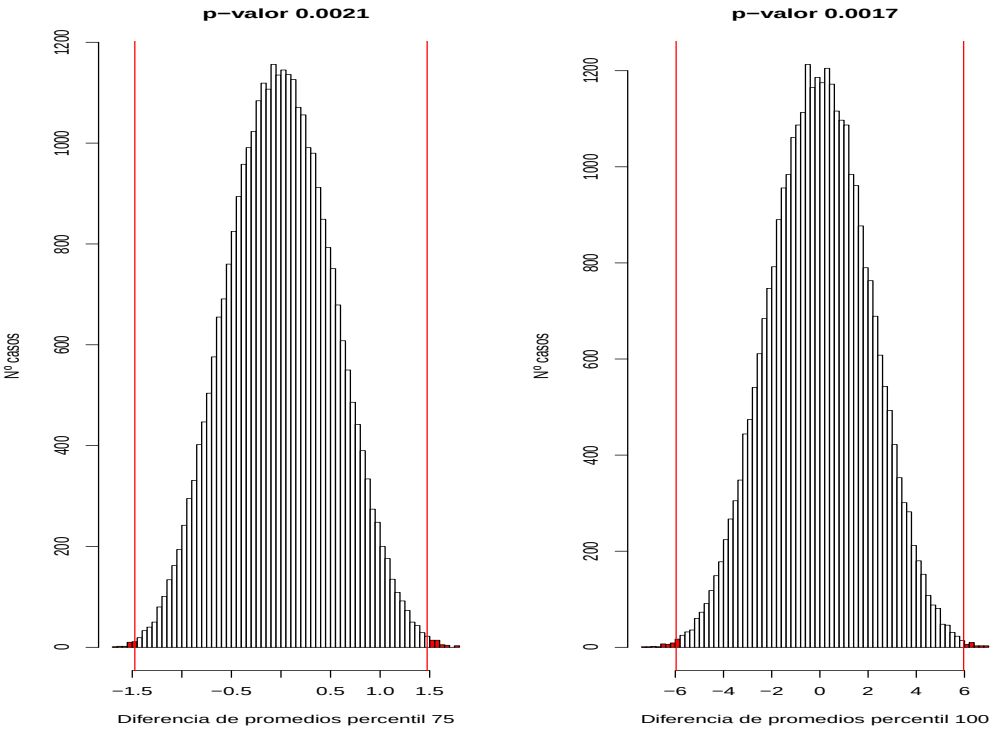
Fig. 9: Grupos resultantes de la tabla (2)

Si para cada sitio hallamos los cuartiles de las velocidades-pico de los eventos, nuevamente observamos una notoria diferencia entre los percentiles 75 y 100 de estos dos grupos, como muestra la figura (10) y corrobora el test de permutaciones (figura (11)).



..

Fig. 10: Percentiles de los grupos robustos de la tabla (2): Nótese la diferencia entre ambos grupos en los percentiles superiores



..

Fig. 11: Distribuciones de aleatorización para las diferencias entre los promedios de los percentiles 75 y 100 en los grupos 1 y 2 de la tabla (2)

6 Asignación de sitios para 45 mts. de altura

Ya sea que usemos la tabla (1) o (2), por un lado quedan sitios sin asignar (8, 12, 30, 44) y sitios que parecen estar en territorio del otro grupo (11, 37, 45, 47).

Reasignaremos los sitios 37, 45 y 47 del Grupo 1 al Grupo 2, el sitio 11 del Grupo 2 al Grupo 1. Los sitios 30 y 44 serán asignados al Grupo 2. Los sitios 8 y 12 se asignarán al Grupo 1.

La figura (12) muestra estos dos grupos en el mapa.

Si calculamos la media de los percentiles 75 y 100 en cada grupo y hacemos la diferencia (Grupo2-Grupo1) que en la muestra observada es positiva, obtenemos los valores para el percentil 75 y para el percentil 100 respectivamente.

Para ver si este valor es algo poco frecuente si ho hubiese diferencia entre los percentiles de ambos grupos, hacemos nuevamente un test de permutaciones.

El p-valor del test es de 0.05 cuando trabajamos a dos colas y de 0.025 cuando trabajamos solamente con la cola izquierda (Grupo2-Grupo1<0).

Es un valor muy pequeño por lo que podemos estar razonablemente seguros que estos dos grupos presentan diferencias significativas en los dos cuartiles más altos. La figura (13) muestra las distribuciones de aleatorización para la diferencia de medias entre grupos en los percentiles 75 y 100. El área roja representa aquellos resultados

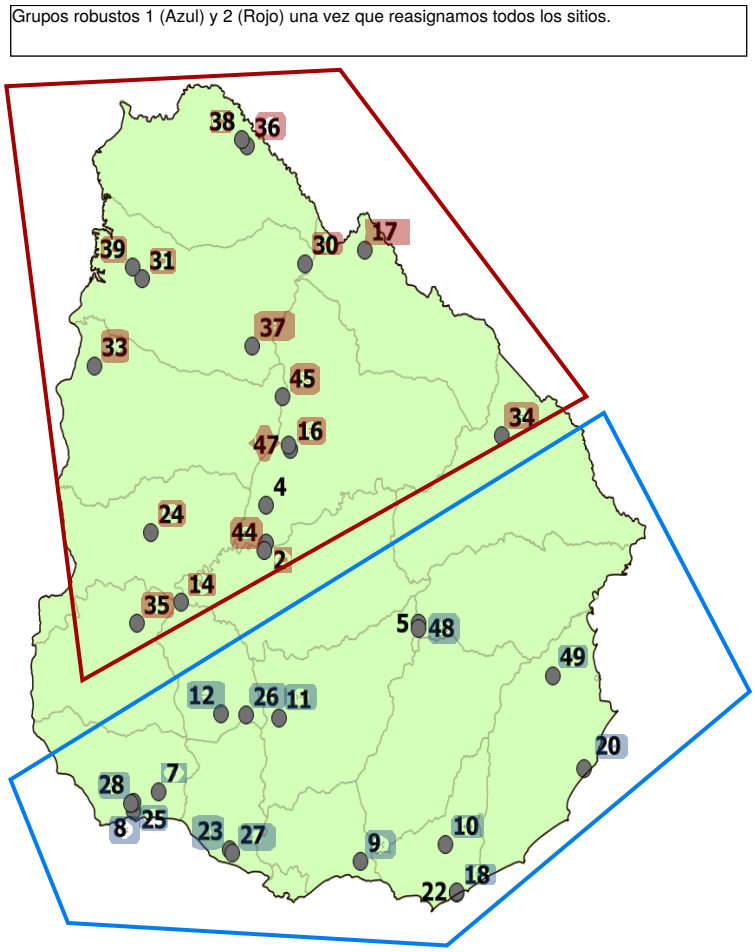
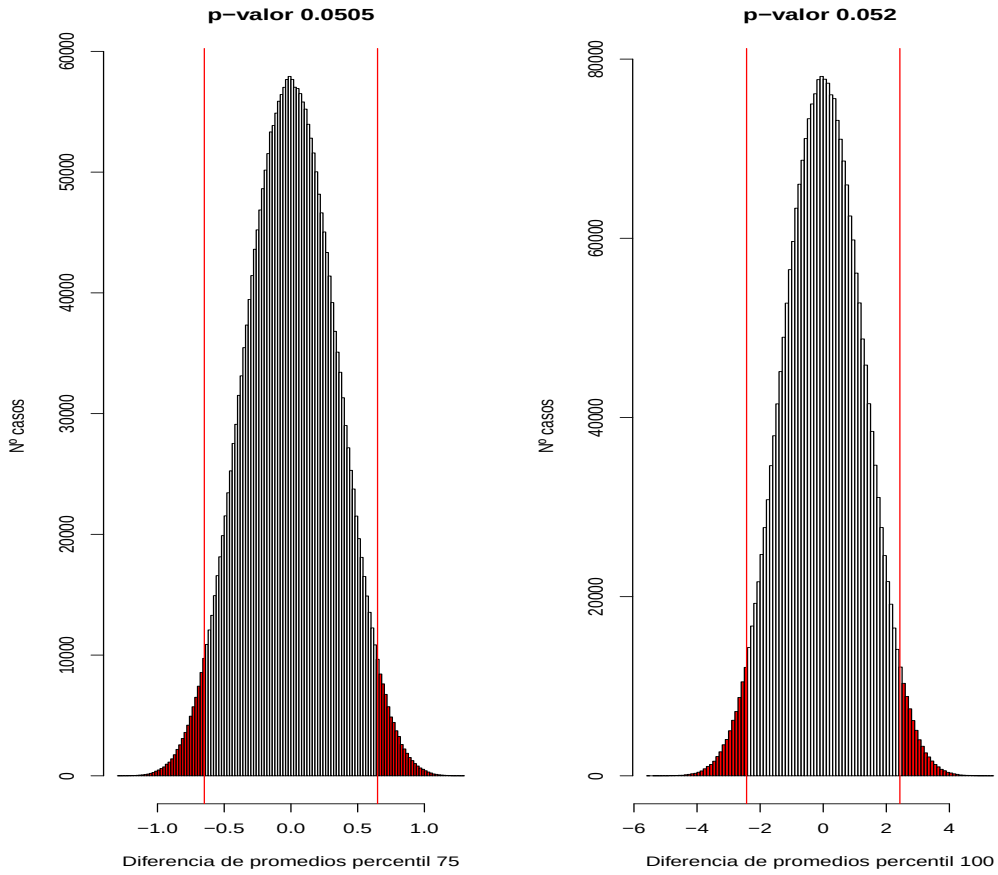


Fig. 12: Grupos construídos a partir de los 2 grupos robustos de la tabla (2) y reasignando sitios



..

Fig. 13: Distribución de aleatorización de las diferencias de medias entre los Grupos 1 y 2 para los percentiles 75 y 100

tanto o más extremos que los obtenidos.

7 Modelando cada grupo para 45 mts. de altura

El siguiente paso es modelar la distribución de los máximos dentro de cada grupo. Para hacer esto juntaremos todos los sitios del Grupo 1 y todos los del Grupo 2.

Como para cada sitio tenemos los máximos semanales, mensuales y trimestrales, tendremos los consiguientes máximos semanales, mensuales y trimestrales del Grupo 1 y del Grupo 2.

Hecho esto ajustaremos las distribuciones Pareto generalizada para cada grupo. Esto permitirá luego calcular tiempos y velocidades de retorno en cada grupo.

7.1 Máximos trimestrales

Trabajaremos con los máximos trimestrales dado que la literatura existente en ingeniería de vientos sugiere que son los reflejan de mejor forma el clima de vientos.

Juntamos los máximos trimestrales de c/u de los sitios del Grupo 1 en una sola muestra, hacemos lo mismo con los sitios del Grupo 2. Recordemos que los grupos se diferenciaban en el último 25% de la muestra, en el cuarto superior de los valores.

Si para cada velocidad v entre 23 y 45 mt/s contamos la cantidad de veces que se excede el umbral v y dividimos luego por el total de datos, tendremos el porcentaje de excedencias respecto del umbral para cada grupo.

La figura (14) muestra que el Grupo 2 muestra porcentajes notoriamente mayores que el Grupo 1 en un rango de umbrales de 28-35 mt/s aproximadamente. Para confirmar estadísticamente que la diferencia es significativa, nuevamente hacemos un test de permutaciones.

Como el Grupo 1 tiene 16 sitios y el Grupo 2 tiene 17 sitios, elegimos al azar 16 sitios de entre los 33 sitios (16+17), juntamos todos los máximos trimestrales y calculamos el porcentaje excedencias para un umbral $v = 31$ mt/s.

Hacemos lo mismo para los 17 sitios restantes y hacemos la diferencia entre el porcentaje del Grupo 2 y el del Grupo 1.

Repetimos esto 1.125.000 veces y contamos cuántas de estas veces la diferencia de porcentajes supera el valor observado. Si dividimos esta cantidad entre el total de veces tendremos el p-valor del test.

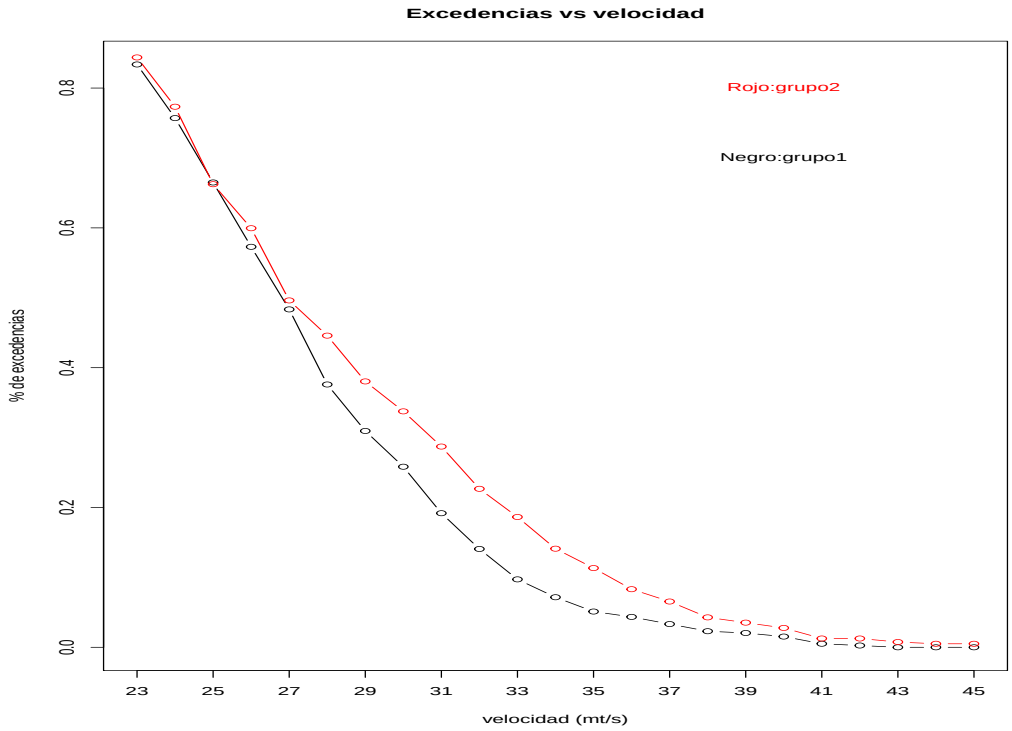
Si no hubiese diferencia entre los grupos, el valor observado debería ser algo que se da con bastante frecuencia, sin embargo los valores tanto o más extremos que el observado representan un 1% de los valores obtenidos, lo que permite afirmar que la diferencia observada en la figura (14) no es producto del azar.

La distribución de aleatorización del test se muestra en la figura (15). Dicho de otra manera, la probabilidad de observar máximos trimestrales por encima de los 31 mt/s es significativamente mayor en el Grupo 2 que en el Grupo 1. Otro tanto puede decirse para umbrales entre 28 y 35 mt/s.

Esto sugiere que para la técnica POT (Peak Over Treshold) que permite modelar la distribución de las excedencias por encima de un umbral, los umbrales a utilizar deben ser no menores a 28 mt/s.

7.2 Peak Over Treshold (POT)

En este enfoque se elige un umbral u adecuado y si las excedencias asociadas son i.id, y se dispone de una cantidad suficiente, se halla el juego de parámetros σ, k (escala y forma respectivamente) de la distribución generalizada de Pareto (GPD) que mejor ajuste al conjunto de excedencias.



..

Fig. 14: Porcentaje de excedencias en función del umbral

La GPD (donde k es el parámetro de forma y σ el de escala) se define como

$$G(x) = 1 - (1 + k(x-u)/\sigma)^{-1/k} \; si \; k \neq 0 \tag{1}$$

$$G(x) = 1 - e^{-(x-u)/\sigma} \; si \; k = 0 \tag{2}$$

El caso $k = 0$ corresponde a la distribución exponencial de parámetro $\lambda = \frac{1}{\sigma}$

Si tomamos como umbral $u = 32$ mt/s, para el Grupo 1 encontramos 55 excedencias en 392 datos (14.03%) y para el Grupo 2, 90 excedencias en 397 datos (22.67%).

Para ver si las excedencias pueden considerarse i.i.d hacemos el test de rangos de Bartels, el test de diferencias de signos, el test de Rachas, test turning point y test de correlación de rangos de Spearman.

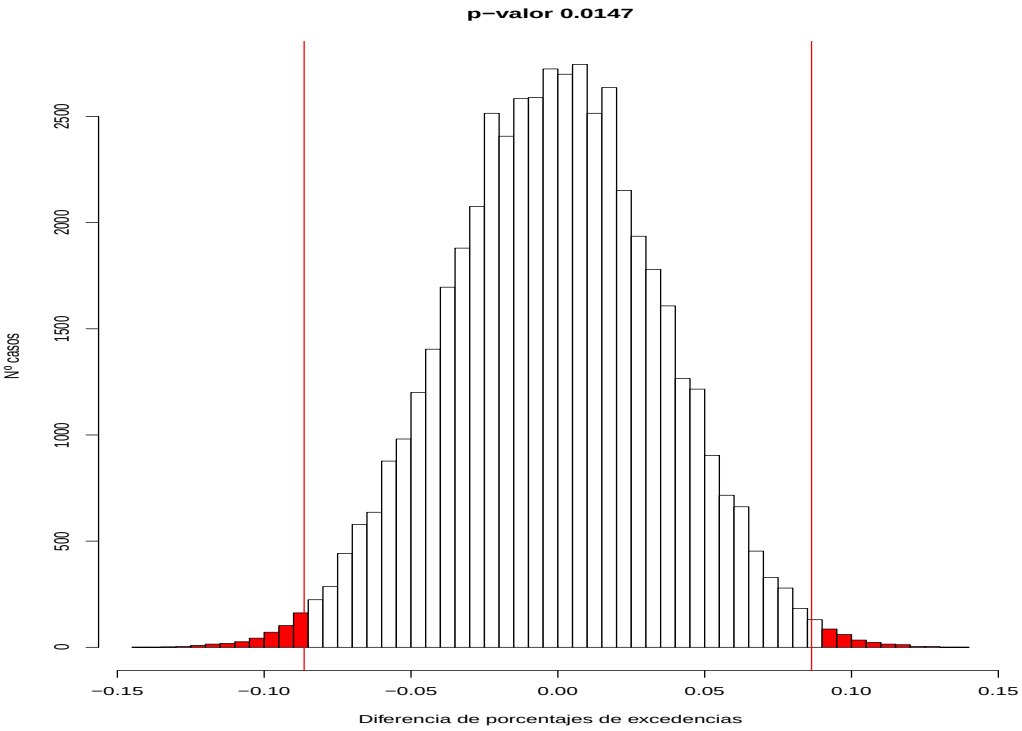
En todos los casos se obtienen p-valores superiores a 0.35 para el Grupo 2 y 0.15 para el Grupo 1 por lo que no rechazamos H_0 (aleatoriedad de la muestra).

Si ajustamos los parámetros de la GPD, obtenemos los IC 095 para los parámetros de forma y escala que caracterizan dicha distribución.

La tabla (3) muestra dichos IC para cada grupo.

La probabilidad de exceder una velocidad v dado que se ha excedido el umbral $u = 32$ mt/s (utilizando las GPD cuyos parámetros se muestran en la tabla (3)), se muestran en la tabla (4).

La figura (16) muestra dichas probabilidades condicionales. Los tiempos de retorno



..

Fig. 15: Distribución de las diferencias de excedencias entre grupos para $v = 32$ mt/s. En rojo se muestran los valores tanto o más extremos que el observado

en cada grupo se muestran en las tablas (5) y (6).

En el Grupo 1 los IC095 son más anchos y con extremos inferiores notoriamente menores que en el Grupo 2.

Sin embargo para el Grupo 1 la tabla (3) muestra que el IC095 para el parámetro de forma incluye el valor 0, por lo que no podemos descartar que la GPD para el Grupo 1 sea una distribución exponencial.

Para tener una confirmación adicional hacemos el test de Lilliefors para exponenciales, cuya H0 es que los datos provienen de una distribución exponencial. El p-valor obtenido para el Grupo 1 es 0.79 por lo que no rechazamos H0.

En el Grupo 2 el IC095 para el parámetro de forma está a la izquierda del cero aunque por poco, por lo que la GPD para dicho grupo podría ser también exponencial.

El test de Lilliefors para el Grupo 2 arroja un p-valor de 0.33, considerablemente más bajo que para el Grupo 1 pero suficientemente grande como para no rechazar H0.

Por tanto no podemos descartar el modelo exponencial para modelar las excedencias en el Grupo 2.

Si ajustamos un modelo exponencial para el Grupo 1 y para el Grupo 2 obtenemos la tabla (7).

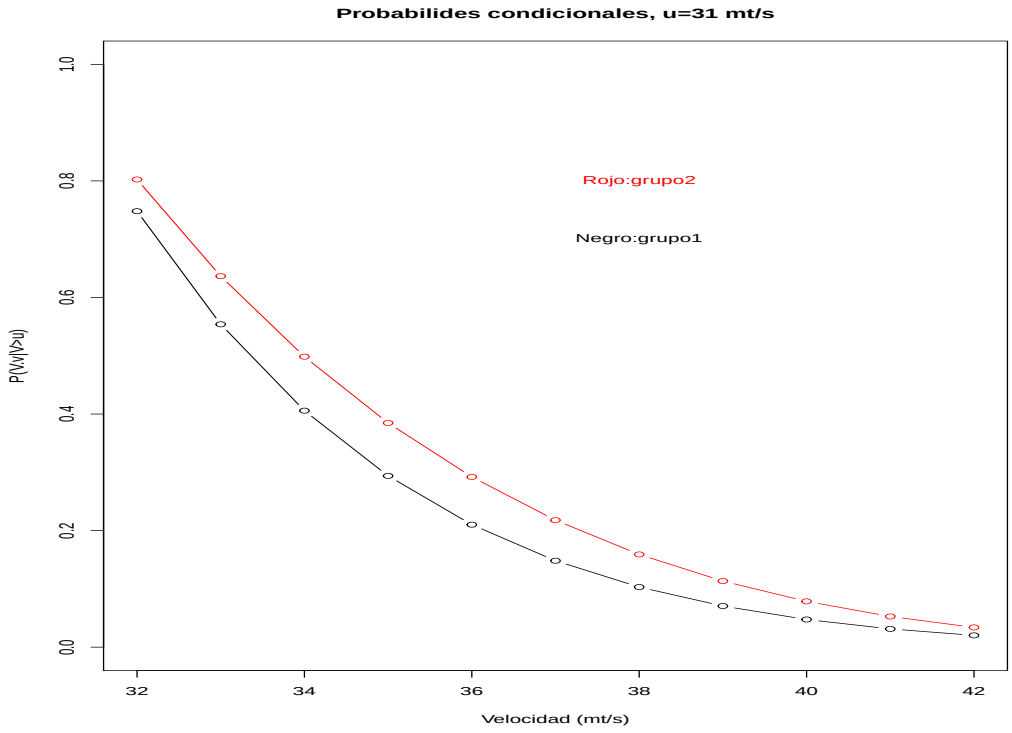
Los niveles de retorno asociados se muestran en la tabla (8). Para el Grupo 2 se

		95% lower CI	Estimate	95% upper CI
Grupo 1	scale	1.60	2.93	4.25
Grupo 1	shape	-0.40	0.00	0.40
Grupo 2	scale	3.33	4.59	5.86
Grupo 2	shape	-0.44	-0.25	-0.06

Tab. 3: Parámetros estimados e IC095 para las GPD en ambos grupos.

v (mt/s)	33	34	35	36	37	38	39	40	41	42
Grupo 1	0.71	0.51	0.36	0.26	0.18	0.13	0.09	0.07	0.05	0.03
Grupo 2	0.80	0.63	0.49	0.37	0.28	0.21	0.15	0.10	0.07	0.04

Tab. 4: $P(V > v|v > u)$, $u = 32$ mt/s



..

Fig. 16: Probabilidades condicionales

observan predicciones mayores que para el Grupo 1.

Vemos que con el modelo exponencial los niveles predichos para el Grupo 1 se diferencia más del Grupo 2.

Niveles de retorno	95% lower CI	Estimate	95% upper CI
5 años	37.00	39.08	41.16
10 años	37.84	41.11	44.39
20 años	38.01	43.14	48.28
50 años	37.33	45.83	54.32
100 años	36.08	47.86	59.63

Tab. 5: Niveles (velocidades) de retorno (en mt/s) para el Grupo 1 según la GPD de la tabla (3)

	95% lower CI	Estimate	95% upper CI
5 años	39.26	41.49	43.71
10 años	40.22	42.91	45.60
20 años	40.83	44.11	47.38
50 años	41.19	45.40	49.61
100 años	41.18	46.20	51.22

Tab. 6: Niveles (velocidades) de retorno (en mt/s) para el Grupo 2 según la GPD de la tabla (3)

8 Modelando eventos Sinópticos y No Sinópticos, 45 mts. de altura

De lo visto hasta ahora surge que en todos los sitios hay una mezcla de los dos tipos de evento y que los eventos no sinópticos suelen ser los que presentan mayores valores de velocidad.

Asimismo los eventos del Grupo 1 son los que presentan mayor proporción de eventos sinópticos, proporción que va decreciendo a favor de los no sinópticos a medida que partiendo del litoral E nos movemos hacia el W-NW.

Esto sugiere que la diferencia en las predicciones y niveles de retorno en ambos grupos se explicarían por el mayor o menor porcentaje de eventos no sinópticos: aquellos sitios con mayor proporción de no sinópticos tendrán mayores velocidades que aquellos donde dicha proporción es menor.

Si volvemos a la introducción, una vez clasificados los eventos de cada sitio con sinópticos o no sinópticos, podemos agrupar todos los eventos sinópticos por un lado y los no sinópticos por otro y aplicar la técnica POT para ver si siguen las mismas distribuciones.

Notemos que ahora no estamos agrupando por sitios (geografía), sino que por tipo de evento, por lo que en el grupo sinópticos irán todos aquellos eventos clasificados como sinópticos, sean del sitio que sean.

Lo mismo para los no sinópticos, por lo que al hablar ahora de dos grupos debemos recordar que no son los mismos grupos de la parte anterior.

8.1 Sinópticos y no sinópticos, son grupos distintos?

Lo primero es comparar la distribución que sigue cada grupo es la misma o no.

La técnica POT requiere contar con una cantidad suficiente de excedencias respecto

	95% lower CI	scale	95% upper CI	
Grupo 1		2.42	3.13	3.84
Grupo 2		2.91	3.67	4.43

Tab. 7: IC095 para el parámetro de las distribuciones exponenciales ajustadas a los Grupos 1 y 2, $u = 32$ mt/s

	95% lower CI	Estimate	95% upper CI	
Grupo 1	5 años	37.61	39.55	41.48
	10 años	39.29	41.72	44.15
	20 años	40.98	43.89	46.81
	50 años	43.20	46.76	50.33
	100 años	44.88	48.94	53.00
Grupo 2	5 años	40.44	42.64	44.84
	10 años	42.46	45.19	47.91
	20 años	44.48	47.73	50.98
	50 años	47.15	51.09	55.04
	100 años	49.17	53.64	58.11

Tab. 8: Niveles (velocidades) de retorno (en mt/s) para los Grupos 1 y 2 según la GPD exponencial de la tabla (7)

de un umbral, y que las mismas sean i.i.d.

Si tomamos como umbral $u = 33$ mt/s para los eventos sinópticos, tenemos 23 datos (0.0003293195 del total) que superan dicho umbral y que además pasan los tests de aleatoriedad anteriormente citados con p-valores superiores a 0.56.

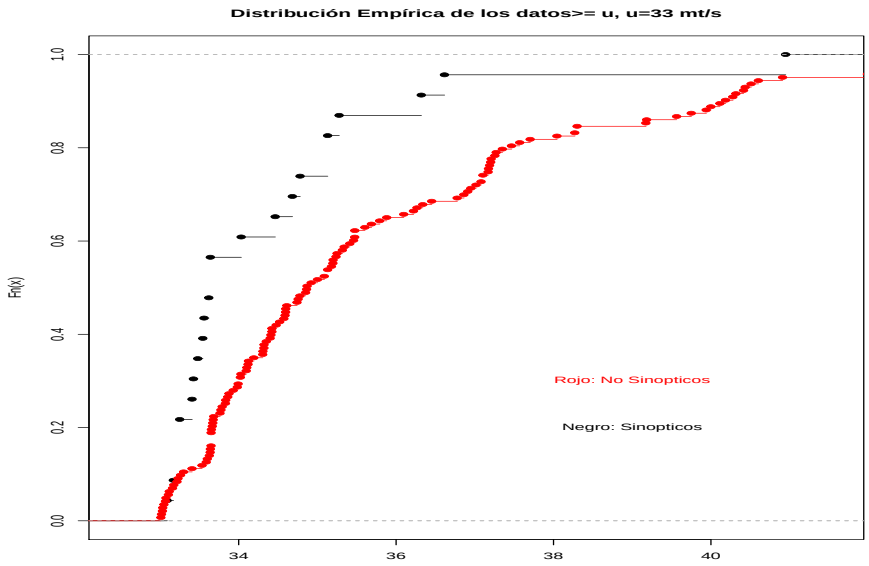
Para el grupo de eventos no sinópticos, tomando $u = 33$ mt/s tenemos 143 datos (0.001657318 del total, el triple que en el caso sinóptico) que superan el umbral y que pasan los tests de aleatoriedad con p-valores superiores a 0.06, por lo que trabajando al nivel 0.95 no rechazamos la hipótesis nula (aleatoriedad de la muestra) .

La figura (17) muestra las funciones de distribución empíricas de las excedencias anteriormente mencionadas.

Es claro que no parecen ser la misma función. Si hacemos el test de comparación de muestras de Kolmogorov-Smirnov (cuya hipótesis nula es que ambas muestras provienen de la misma distribución), obtenemos un p-valor de 0.001944741, rechazando enfáticamente H0.

Si ajustamos la función de densidad de probabilidad para cada juego de datos, obtenemos la figura (18).

La misma muestra que es poco creíble que ambas muestras provengan de la misma ley estadística.



..

Fig. 17: Funciones de distribución de los datos que superan el umbral.

8.2 Calculando los parámetros de las GPD para cada grupo

Una vez que confirmamos que cada grupo tiene una ley distinta, procedemos a ajustar la GPD para cada grupo.

Los parámetros de la GPD para cada grupo junto con sus IC095 se muestran en la tabla(9).

		95% lower CI	Estimate	95% upper CI
sinópticos	scale	0.43	1.16	1.90
sinópticos	shape	-0.30	0.18	0.67
no sinópticos	scale	2.34	3.12	3.89
no sinópticos	shape	-0.28	-0.09	0.10

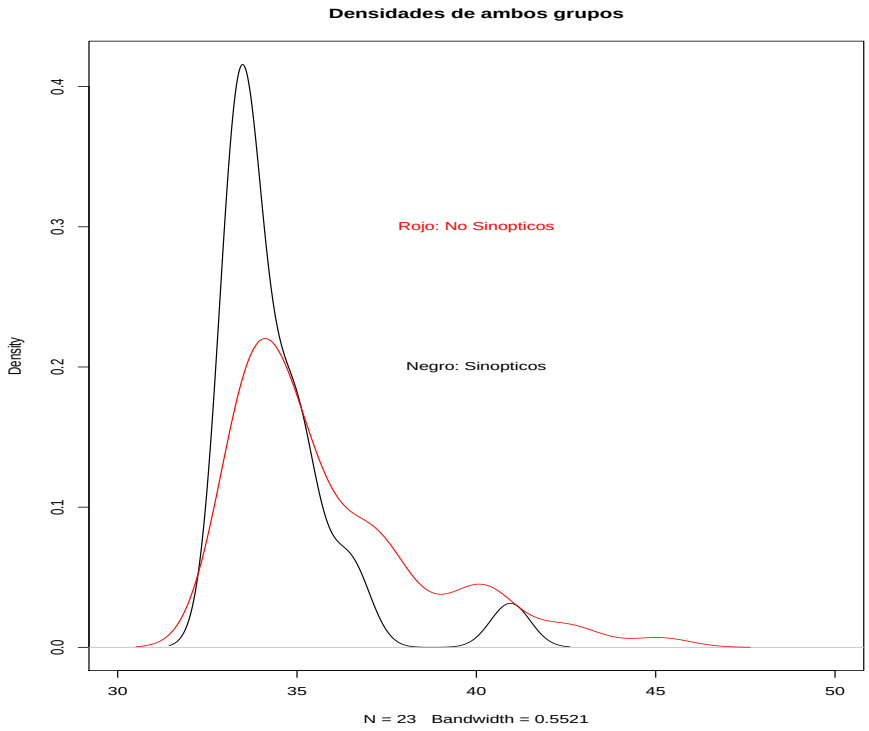
Tab. 9: Estimación de los parámetros de la GPD en cada grupo. Los IC para el parámetro de forma incluyen el 0

Vemos que en ambos casos el IC para el parámetro de forma no permite excluír el valor 0, en el caso del grupo no sinóptico el IC es más angosto porque se cuenta con más datos.

Por tanto ajustaremos un modelo exponencial a cada grupo.

Las estimaciones del parámetro de escala junto con los IC095 se muestran en la tabla (10).

Para tener una confirmación adicional, hacemos el test de Lilliefors a los datos (excedencias) de ambos grupos. Los p-valores para los grupos Sinóptico y No Sinóptico son 0.090 y 0.289 respectivamente.



..

Fig. 18: Densidades de las excedencias para los grupos sinópticos y no sinópticos

Es de notar que los IC095 para el parámetro de forma no se solapan, están separados.

8.3 Comparación de los modelos ajustados con los datos

Para validar las GPD halladas (los dos modelos exponenciales) generamos 1000 juegos de 1000 datos, c/u generado por la GPD correspondiente. Las figuras (19) y (20) muestran la densidad de los datos (en negro) y la de una de las muestras simuladas (en rojo).

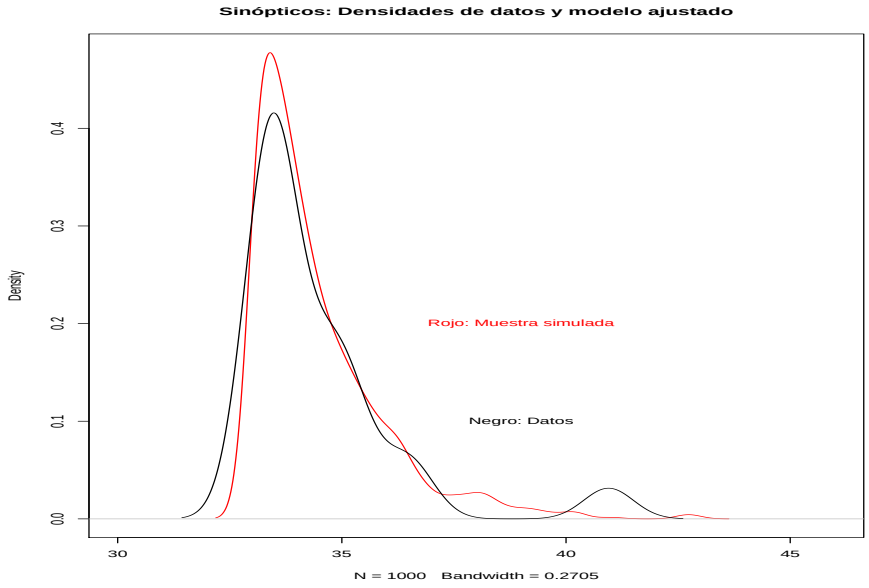
Para cada juego de datos hacemos el test de comparación de muestras de Kolmogorov-Smirnov, comparando las muestras sintéticas con los datos.

La hipótesis nula es que ambas muestras provienen de la misma distribución.

Tendremos 1000 p-valores para el caso sinóptico y otro tanto para el caso no sinóptico.

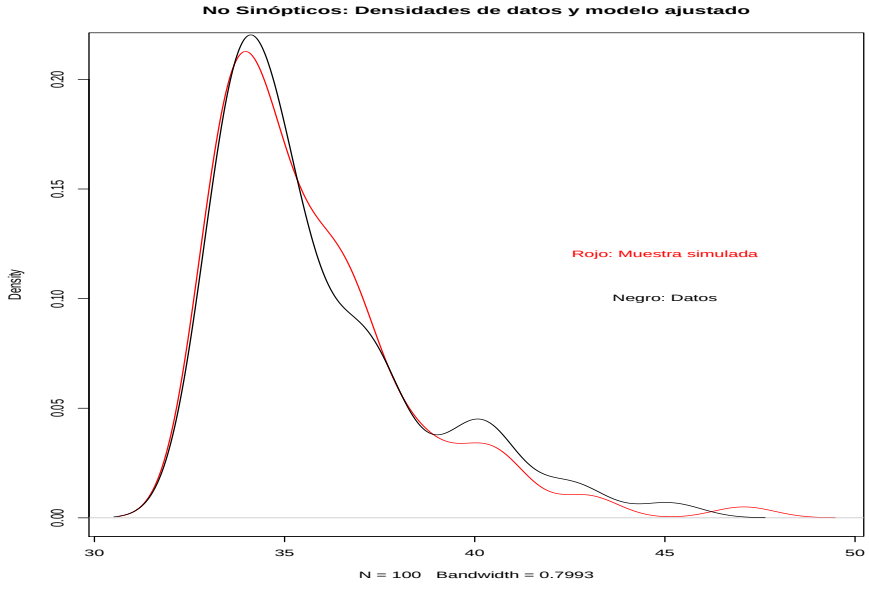
Para el grupo sinóptico, los p-valores obtenidos son mayores a 0.11 con mediana 0.32 y media 0.32, para el no sinóptico mayores a 0.13 con mediana 0.59 y media 0.57, por lo que no rechazamos H0: las distribuciones ajustadas a cada grupo generan muestras creíbles.

Las figuras (21) y (22) muestran la distribución de estos 1000 p-valores para cada grupo.



..

Fig. 19: Grupo Sinóptico: Densidad de los datos y de una de las muestras generadas por el modelo exponencial hallado

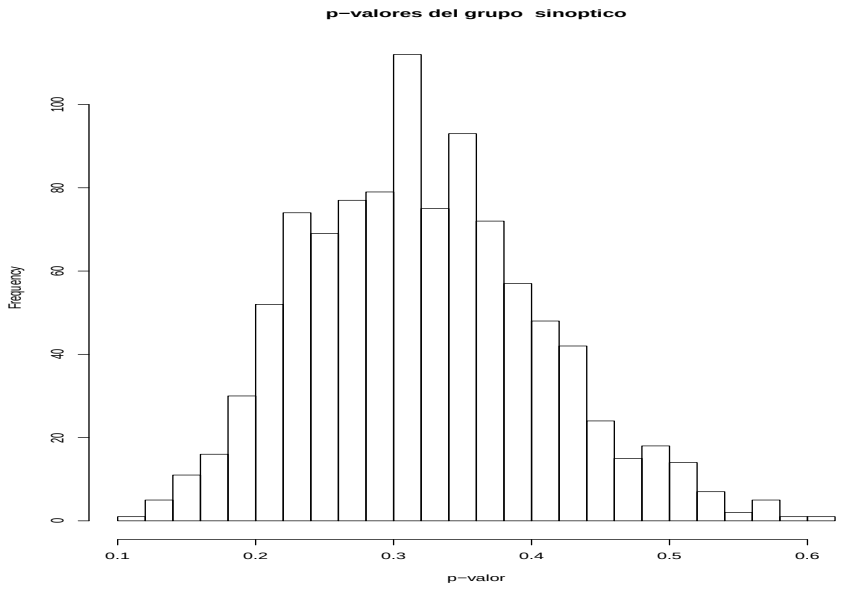


..

Fig. 20: Grupo No Sinóptico: Densidad de los datos y de una de las muestras generadas por el modelo exponencial hallado

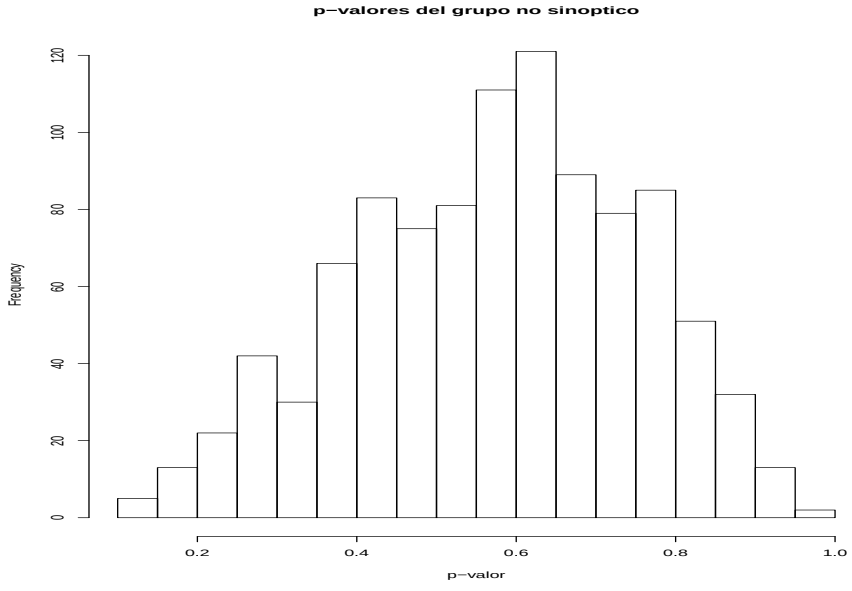
	95% lower CI	scale	95% upper CI
sinópticos	0.84	1.42	2.00
no sinópticos	2.39	2.85	3.32

Tab. 10: IC095 para los paráemtros de forma en ambos modelos exponenciales



..

Fig. 21: Distribución de los 1000 p-valores al comparar los datos con las muestras simuladas, grupo Sinóptico



..

Fig. 22: Distribución de los 1000 p-valores al comparar los datos con las muestras simuladas, grupo No Sinóptico